L'imputazione dei dati mancanti: l'effetto sui parametri di un Extended Logistic Rasch Model

Missing data imputation: the effect on Extended Logistic Rasch Modele parameters

Boscaino G. e Sulis I.

DSSM Working Papers - n. 2008.3
Palermo - 11 marzo 2008

Giovanni Boscaino
Dipartimento di Scienze Statistiche e Matematiche 'S. Vianelli'
Università degli Studi di Palermo

Isabella Sulis
Dipartimento di Ricerche Economiche e Sociali
Università di Cagliari

Corresponding author:

Giovanni Boscaino
Tel: +390916626334
e-mail: gioboscaino@unipa.it
web site: www.unipa.it/gioboscaino

L'imputazione dei dati mancanti: l'effetto sui parametri di un Extended Logistic Rasch Model[1]

Giovanni Boscaino[2] e Isabella Sulis[3]

Introduzione

Il problema dei dati mancanti è abbastanza comune nella ricerca empirica, specialmente nelle scienze sociali in cui il tentativo di misurazione di quantità non direttamente osservabili (variabili latenti) avviene attraverso la somministrazione di test o questionari costituiti da più *item*. I modelli statistici finalizzati alla soluzione di tale problema richiedono, in genere, un elevato numero di osservazioni per ogni unità coinvolta nell'analisi. In un contesto multivariato il problema si amplifica, poiché nel modello sono considerati più *item* per ciascuna osservazione: la probabilità, quindi, di avere almeno un dato mancante non è irrilevante ed è, inoltre, crescente al crescere del numero di *item*. Dopo una breve panoramica di alcuni dei principali approcci ai dati mancanti, il lavoro pone l'attenzione sul metodo della Multiple Imputation, valutandone i vantaggi tramite il confronto con altri tre approcci noti e largamente usati in letteratura. L'esempio che viene riportato si focalizza sul confronto dell'efficienza delle stime dei parametri di posizione degli *item* quando si applica l'Extended Logistic Model (ELM) di Rasch. L'ambito applicativo a cui si vuole fare qui riferimento è quello della valutazione della didattica universitaria, da anni adottata da tutti gli Atenei pubblici italiani.

Keywords: Multiple Imputation, Rasch Model, Valutazione Qualità della Didattica, 'Taratura' del questionario

1 – *Il problema delle mancate risposte nelle indagini statistiche*

Nell'ambito delle scienze sociali la somministrazione di questionari è una delle tecniche più diffuse per la raccolta di dati e informazioni. Uno dei primi problemi che un ricercatore si trova ad affrontare, in fase di analisi dei risultati, è quello di un dataset incompleto e con errori. Questo accade generalmente perché chi compila il questionario non ne interpreta correttamente la struttura, commette accidentalmente qualche errore nel fornire le risposte, non vuole deliberatamente rispondere ad alcune domande, oppure a causa di un errore dello strumento di codifica, che dal supporto cartaceo (il più utilizzato nella maggior parte dei casi) deve trasferire i dati su supporto informatico, o di chi invece si occupa del data entry (Knapp, 1998).

Non esiste in letteratura un'unica tecnica o una metodologia di approccio al problema di come tenere sotto controllo l'effetto dei dati mancanti: ogni situazione è un caso a sé (Rubin, 1976; Schafer and Graham, 2002; Tsikriktsis, 2005). In generale, è sempre consigliabile saggiare lo strumento di rilevazione con indagini pilota in modo da studiarne punti di forza e di debolezza, in modo da intervenire su questi ultimi e prevenire la presenza di risposte omesse (Stone, 2001). Nel momento in cui, nonostante tutti gli accorgimenti, il problema continua a presentarsi, la quantità e la distribuzione dei *missing data*, la struttura dei dati e la natura delle variabili coinvolte, saranno l'unica indicazione in base alla quale prendere decisioni. La trattazione delle possibili soluzioni note in campo

[1] Il lavoro presente è frutto delle comuni riflessioni e ricerche degli autori; in particolare G. Boscaino ha curato il paragrafo 3, I. Sulis i paragrafi 1 e 2. Si ringraziano, inoltre, la prof.ssa Vincenza Capursi e il dott. Vito Muggeo per i preziosi consigli e suggerimenti.

[2] Giovanni Boscaino è titolare di assegno di ricerca, presso il Dipartimento Scienze statistiche e Matematiche 'S. Vianelli', Università degli Studi di Palermo.

[3] Isabella Sulis è titolare di assegno di ricerca presso il Dipartimento di Ricerche Economiche e Sociali, Università di Cagliari.

statistico esula, però, dagli scopi di questo lavoro; pertanto, per una rassegna esaustiva dell'argomento si rimanda alla letteratura specializzata (Rubin, 1988; Schafer, 1997; Little e Rubin, 2002).

2 – *Analisi con dati mancanti: implicazioni e tecniche adottate*

2.1 Tipologie di dati mancanti

La presenza di dati mancanti può influire sulle performance degli stimatori e pertanto può condurre a risultati inferenziali non corretti. In questo contesto è determinante stabilire se il meccanismo che ha generato i valori mancanti è di tipo casuale o meno, nonché analizzare le possibili relazioni tra i valori mancanti e i dati effettivamente rilevati.

Data la matrice dei dati osservati $\boldsymbol{X}$, il cui elemento generico è x_{ij}, si definisce una matrice $\boldsymbol{R}$ i cui vettori colonna sono delle variabili indicatrici di evento in base alle quali il suo generico elemento r_{ij} è pari a 1 se x_{ij} è un dato mancante e r_{ij}=0 se x_{ij} è osservato. Tale matrice, nota come matrice di *missingness* (Rubin, 1976), è trattata come un insieme di variabili casuali legate da una distribuzione di probabilità congiunta. L'intera matrice dei dati può essere partizionata nelle due sottomatrici di dati osservati ($\boldsymbol{X}_{obs}$) e dati mancanti ($\boldsymbol{X}_{miss}$): $\boldsymbol{X}=(\boldsymbol{X}_{obs},\boldsymbol{X}_{miss})$. In base alla distribuzione di probabilità di $\boldsymbol{R}$ e alla sua relazione con la matrice dei dati, Rubin (1976) identifica tre tipologie di dati mancanti:

- *Missing Completely at Random* (MCAR), se la presenza del dato mancante è assolutamente dovuta al caso e la probabilità di osservare i dati mancanti per un *item* è indipendente dalle altre risposte date a quello e a gli altri *item,* ma dipende solamente dal parametro φ che caratterizza la distribuzione di R ($P(\boldsymbol{R}|\boldsymbol{X}_{com,,}\varphi)=P(\boldsymbol{R}|\varphi)$);
- *Missing at Random* (MAR), se la probabilità di osservare un'unità mancante dipende solo dai dati osservati ($P(\boldsymbol{R}|\boldsymbol{X}_{com,}\varphi)=P(\boldsymbol{R}|\boldsymbol{X}_{obs,}\varphi)$);
- *Missing Not At Random* (MNAR), se la distribuzione di probabilità associate a $\boldsymbol{R}$ dipende sia dai dati osservati che dalla distribuzione dei dati mancanti ($P(\boldsymbol{R}|\boldsymbol{X}_{com,}\varphi)=P(\boldsymbol{R}|\boldsymbol{X}_{obs},\boldsymbol{X}_{miss,}\varphi)$[4].

L'assunzione che il meccanismo di generazione dei dati mancanti sia MAR implica che i valori mancanti sono prevedibili in base alle risposte osservate nelle altre variabili; al contrario, quando i dati mancanti sono MNAR, l'informazione contenuta nel dataset non è sufficiente per predire ciò che non è stato osservato.

2.2 Alcune considerazioni sulle tecniche di gestione dei dati mancanti

Le tecniche di gestione dei dati mancanti si possono distinguere in tre gruppi: l'eliminazione; la sostituzione (imputazione) dei *missing data* con valori stimati; la modellazione della distribuzione dei dati mancanti.

Una delle prime tecniche di gestione dei *missing data* (da un punto di vista storico) è anche la più immediata: l'eliminazione dall'analisi di qualsiasi unità sia stata parzialmente osservata (*Complete Case Analysis*, CCA). Tale tecnica è, spesso, adottata automaticamente dalla maggior parte dei software statistici, sia in un contesto univariato che multivariato, sotto l'assunzione che il meccanismo che ha generato i dati mancanti sia casuale e che, quindi, queste osservazioni possano essere definite *Missing Completely At Random*. Quando questa assunzione non è soddisfatta, la pratica di procedere con una analisi dei soli casi completi può generare gravi ripercussioni in termini di risul-

[4] Il termine "random" adottato da Rubin non deve essere interpretato in un'ottica prettamente statistica: non si sta facendo riferimento ad un processo probabilistico, anche perché tutte e tre le configurazioni di dai mancanti presuppongono una distribuzione di probabilità per $\boldsymbol{R}$. Piuttosto con tale termine si vuole fare riferimento ad un processo non prevedibile e estraneo alle variabili in oggetto di studio (Schafer e Graham, 2002).

tati inferenziali, specialmente quando la probabilità di non rispondere ad un *item* è connessa alle caratteristiche delle unità o ad altre variabili che sono osservate nell'indagine. Le conseguenze dirette sono quantificabili in termini di distorsione e/o inefficienza delle stime, a seconda che i dati siano MAR, MCAR o MNAR. Quando i dati sono MNAR o MAR le sole osservazioni complete, infatti, non possono essere considerate un campione casuale estratto dalla matrice dei dati originali. Inoltre, anche quando si può asserire che molto verosimilmente la distribuzione dei dati mancanti è completamente casuale (MCAR), la sola riduzione della numerosità campionaria è sufficiente a provocare una perdita di efficienza.

Per riassumere, le conseguenze causate, in termini di correttezza e precisione delle stime, dalla scelta di procedere con un'analisi delle sole unità totalmente osservate, sono strettamente connesse alla riduzione della numerosità campionaria, a quanto la struttura dei dati completi differisce da quella delle unità parzialmente osservate, nonché, alla natura dei parametri che sono oggetto di stima nelle analisi (Little e Rubin, 2002). Da un punto di vista operativo, se le osservazioni omesse rappresentano una quota marginale della numerosità campionaria[5] e il profilo delle caratteristiche individuali non presenta dei pattern particolari, procedere con un'analisi dei soli dati completi è sicuramente una procedura semplice e ragionevole. Nella pratica, tuttavia, le conseguenze dell'omissione sono difficilmente accertabili sulla base delle informazioni che si hanno a priori.

Un'altra soluzione, frequentemente usata per evitare lo "spreco" di dati che si ha quando si adotta una CCA, è procedere con un'analisi dei soli casi disponibili (*Available Case Analysis,* ACA), utilizzando differenti sottoinsiemi di unità per la stima di diversi parametri. Operando in questo modo, la numerosità del campione varia nel processo di stima a seconda del pattern di osservazioni mancanti, rendendo, in alcuni casi, non semplice il calcolo dei relativi standard error. Inoltre, diversi studi mostrano che l'adozione di questa tecnica è sconsigliabile per la stima dei parametri di un modello regressivo (Haitovsky, 1968; Glasser, 1966; Little, 1992), soprattutto quando la frazione di dati mancanti non è moderata, le variabili interessate sono fortemente correlate e il modello oggetto di analisi non è correttamente specificato.

Nell'ambito dell'*Item Response Theory* la scelta di procedere con un'analisi delle osservazioni disponibili non è ammessa: questa classe di modelli richiede come requisito fondamentale che il set di misure ripetute appartenenti alla stessa unità sia completamente osservato, in quanto i totali di riga e di colonna costituiscono statistiche sufficienti per la stima dei parametri chiave del modello. Questa condizione implica che ogni qualvolta si osservano dati mancanti, si dovrebbe verificare che la condizione di "ignorabilità" sia soddisfatta[6].

Le tecniche che tengono conto della presenza di dati mancanti durante la fase di stima dei parametri (utilizzando procedure iterative quali ad esempio l'algoritmo *Expectation–Maximization*, EM) basandosi sulla modellazione della funzione di verosimiglianza congiunta delle unità totalmente e parzialmente osservate, richiedono per la loro applicabilità che ipotesi forti (specialmente quando le variabili coinvolte sono misurate su diverse scale) siano avanzate sulla distribuzione congiunta delle variabili (ad esempio multinormalità).

Da un punto di vista operativo, quando il dataset contiene informazioni che possono essere utilizzate a fini predittivi, una pratica più attraente e che, inoltre, permette maggiore flessibilità è sicuramente quella di sostituire i valori non osservati con valori plausibili generati sulla base di un modello probabilistico. Il valore imputato per la data unità dovrebbe essere considerato il più plausibile dato il pattern di risposte osservate per l'unità *i* e per le altre osservazioni nel dataset. In questo modo si utilizza tutta l'informazione disponibile sulle unità parzialmente osservate e il dataset con valori imputati può essere analizzato con i metodi statistici standard. Il principale vantaggio de-

[5] Schafer (1997) suggerisce che i *missing data* non dovrebbero essere ignorati se presenti in percentuale superiore al 5%.

[6] Questa condizione è difficilmente verificabile a priori. Nella pratica una prima indicazione può essere fornita dalla presenza di pattern particolari nella distribuzione dei dati mancanti relativamente alle altre variabili osservate nel dataset (ad esempio le caratteristiche dei rispondenti) .

rivante dall'imputazione delle unità non osservate è il mantenimento dell'ampiezza campionaria che, quindi, si riflette sulla potenza dei test statistici.

Un forte contributo alla tecniche di gestione dei dati mancanti si è sviluppato in campo bayesiano, dove interessanti prospettive sono state aperte dall'adozione di metodi di Monte Carlo Markov Chain, MCMC (Gibbs sampling, data augmentation, l'algoritmo di Metropolis-Hastings) (Schafer, 1997). Questi metodi, tuttavia, come tutte le tecniche di impronta bayesiana, richiedono la formulazione di ipotesi sulla distribuzione a priori dei parametri al fine di campionare i valori mancanti dalla distribuzione a posteriori di X_{miss}. Il grande investimento in termini computazionali che queste procedure richiedono, le difficoltà di implementazione connesse e, soprattutto, la scarsa disponibilità di software con procedure implementate, ne hanno fino a qualche anno fa limitato l'utilizzo nell'ambito della ricerca applicata.

Le differenze tra i vari metodi si riducono all'aumento dell'ampiezza campionaria, alla riduzione del numero di *missing value* e del numero di variabili interessate da dati mancanti (Raymond, 1986).

2.3 Le tecniche considerate

Nel dataset sulla valutazione della qualità dei corsi universitari si è supposta una struttura MAR per i dati mancanti. Sotto questa ipotesi si è voluto valutare gli effetti di quattro procedure di gestione dei *missing data*: (a) l'analisi dei soli dati completi, (b) la sostituzione dei dati mancanti con quelli modali dell'*item*; (c) la sostituzione dei valori mancanti con i valori predetti sulla base di un modello di regressione; (d) la sostituzione dei dati mancanti di una variabile con estrazioni casuali dalle rispettive distribuzioni condizionate (Raghunatan *et al.*, 2001, Little and Rubin 2002). Nell'approccio (c), data la struttura multivariata dei dati mancanti, si è utilizzato un sistema di equazioni di regressione parametriche sequenziali, mentre in (d) si è proceduto con una *Multiple Imputation Analysis* (Rubin, 1988), sfruttando un sistema di regressioni stocastiche sequenziali (Raghunatan *et al.*, 2001).

Si può facilmente intuire come la scelta di sostituire i valori mancanti con i valori modali della distribuzione è un metodo che può severamente sottostimare le misure di dispersione. Questa pratica, ampiamente utilizzata in passato, è generalmente sconsigliata quando la frazione di unità non osservate non è ignorabile e quando i parametri oggetto di stima coinvolgono le relazioni tra più variabili del dataset. Tuttavia, data la rapidità e la semplicità di applicazione del metodo e, soprattutto, il fatto che le conseguenze sulle stime dei parametri sono impercettibili quando la frazione di unità non osservate è minima, si è deciso di inserirla tra le tecniche di gestione dei dati mancanti che sono state poste a confronto.

Nel terzo approccio i valori mancanti sono stati sostituiti con i valori attesi della variabile, condizionatamente alle risposte osservate negli altri *item*. A tal scopo, per la predizione dei valori mancanti in ciascuno degli *item* affetto da *missingness*, è stato utilizzato un sistema di regressioni sequenziali parametriche (Raghunatan *et al.* 2001). In tal modo, il processo di imputazione in una struttura multivariata è stato ridotto a una serie di singoli modelli regressivi che possono essere facilmente specificati a seconda della natura della variabile coinvolta.

Il quarto approccio, introdotto da Rubin nel 1987, prevede la sostituzione di ciascun valore mancante $x_{i.miss}$ con m estrazioni casuali da una distribuzione ritenuta plausibile per $x_{i.miss}$. Sostituendo in tutti gli *item* del dataset i valori generati nella prima estrazione, poi quelli generati nella seconda, e così via sino alla *m-esima*, si ottengono m matrici dei dati complete e che possono essere analizzate separatamente con l'utilizzo di strumenti statistici standard. Nella fase successiva, i valori delle stime dei parametri e delle relative misure di incertezza ottenuti su ogni dataset sono sintetizzati in un'unica misura, in modo da avere un singolo risultato inferenziale che tenga espressamente in considerazione la variabilità osservata nelle stime (Rubin, 1987). Questo metodo di imputazione stocastico è considerato superiore rispetto ai precedenti. La pratica di imputare più di un plausibile valore per ogni *item* affetto da *missingness* permette di specificare l'intera distribuzione del valore mancante, nonché di tenere in considerazione, nella fase di analisi dei dati, l'incertezza sul vero va-

lore del parametro stimato. Pertanto, differentemente dagli altri metodi citati, sviluppando una *Multiple Imputation Analysis* i valori imputati non sono trattati, in fase di analisi come reali.

In un contesto con struttura multivariata dei dati mancanti, dove in tutti gli *item* si registrano valori non osservati e il *pattern* dei dati mancanti è non monotono (le variabili non possono essere ordinate al fine di evitare che vi siano osservazioni mancanti tra i predittori), generare i valori plausibili con l'utilizzo di modelli di tipo regressivo non è di semplice implementazione.

Le procedure di imputazione adottate (c) e (d) superano questo problema utilizzando una versione modificata[7] del *Sequential Regression Imputation Method* proposto da *Raghunathal et al.* (2001). La versione che è stata implementata[8] specifica inizialmente un modello di regressione per ogni *item* coinvolto nell'analisi e poi procede iterando più volte il sistema di equazioni sequenziali, specificatamente:

1. per ogni variabile *Y* affetta da osservazioni mancanti viene individuato il miglior predittore lineare;

2. per ogni variabile *Y* vengono i prima 'imputati i valori mancanti per cui sono osservati tutti i *k* regressori specificati nel predittore lineare, successivamente, quei valori di *Y* per cui sono osservati solamente *k-1* dei regressori inizialmente selezionati e così via sino a che tutte le unità mancanti per cui è osservabile almeno un regressore vengono sostituite con un valore plausibile;

3. nel metodo (c) i valori mancanti per ogni *item Y* sono predetti utilizzando diversi predittori , a seconda del pattern di unità non osservate tra i regressori del modello;

4. nel metodo (d) i valori mancanti sono generati come estrazioni casuali da diverse distribuzioni condizionate, i cui parametri, stimati utilizzando modelli regressivi parametrici, dipendono dal pattern di unità non osservate tra i regressori del modello;

5. la procedura è iterata più volte, per aggiornare i coefficienti dei parametri dei modelli di regressione – metodo (c) – o delle distribuzioni da cui i dati sono generati – metodo (d) – in funzione dell'ultima informazione disponibile sui valori imputati;

6. ad ogni iterazione le variabili imputate nel passo precedente sono utilizzate come predittori, fino a che tutte le unità non osservate in un *item s*ono imputate dalla distribuzione della variabile *Y* condizionatamente al set di migliori predittori;

7. la procedura viene applicata a tutte le variabili affette da valori mancanti.

Nel dataset imputato con il metodo (c), le unità mancanti sono state sostituite a ogni iterazione con i valori predetti dal modello di regressione con il livello più elevato di accuratezza nella predizione. Nell'applicazione proposta nel paragrafo successivo è stato sempre selezionato un modello di regressione logistica multinomiale, dato che gli *item* sono misurati su scala Likert a quattro categorie.

---

[7] Per approfondimenti sulla procedura e la sua validazione si rimanda a Sulis I. , 2007.. Nel lavoro si mira ad accertare, attraverso il ricorso a metodi di simulazione, se la procedura implementata soddisfa i requisiti dell'"*accuracy in distribution*" e "*accuracy in estimation*". Inoltre, attraverso uno studio comparativo tra la stima dei valori dei parametri in un dataset originario e nello stesso dataset con dati prima simulati come mancanti e poi imputati con il metodo proposto, si è verificata l'accuratezza e l'affidabilità della procedura di imputazione multipla al variare della frazione di dati mancanti nel data set..

[8] La procedura è stata implementata con il supporto del software R e della relativa libreria *nnet* per la stima dei parametri di un modello di regressione logistica multinomiale.

Nei dataset imputati con l'approccio stocastico (d), sono stati generati *m* valori plausibili per ciascun valore mancante, attraverso generazioni casuali da una distribuzione multinomiale con i seguenti parametri:

$\tilde{Y}_i \sim Multinomiale(\hat{\pi}_1,...,\hat{\pi}_S)$

dove $\hat{\pi}_s = \dfrac{\exp(\alpha_s + \beta_s' x)}{1+\sum_{h=1}^{S-1}\exp(\alpha_h + \beta_h' x)}$.

Gli *m* dataset generati alla fine del processo di imputazione sono stati analizzati mediante la *Multiple Imputation Analysis* (MI). Per ogni parametro θ la stima finale $\bar{\hat{\theta}}$ è stata ottenuta come la media delle stime $\hat{\theta}$ osservate negli *m* dataset

$$\bar{\hat{\theta}} = m^{-1}\sum_{i=1}^{m}\hat{\theta}_i . \qquad [1]$$

La varianza (*T*) della stima è il risultato della combinazione di due fonti di variabilità: la media della varianza delle stime all'interno di ciascun dataset, chiamata anche varianza *within*

$$W = m^{-1}\sum_{i=1}^{m}U^i , \qquad [2]$$

dove $\sqrt{U^i}$ è lo standard error di $\hat{\theta}_i$, e la variabilità tra le *m* stime $\hat{\theta}_i$

$$B = (m-1)^{-1}\sum_{i=1}^{m}(\hat{\theta}_i - \bar{\hat{\theta}})^2 , \qquad [3]$$

che è chiamata varianza *between*. La varianza totale risulta dalla combinazione lineare delle due componenti[9]

$$var(\bar{\hat{\theta}}) = T = W + \left(1+m^{-1}\right)B . \qquad [4]$$

La componente *within* è considerata la varianza che si osserverebbe in assenza di osservazioni mancanti nel dataset, mentre la varianza *between* racchiude l'incertezza sul vero valore del parametro perchè assume valore 0 quando la frazione di dati mancanti è nulla o, prossima allo zero, quando le stime di θ nei dataset imputati sono simili: minore è la variabilità tra le stime del parametro θ, minore è l'informazione contenuta nelle unità mancanti. Rubin (1987) mostra che solitamente un numero limitato *m* di imputazioni è sufficiente per avere validi risultati inferenziali, anche quando la percentuale di valori mancanti è moderatamente elevata. Pur considerando la MI un metodo di Monte Carlo, in generale questa tecnica non necessita di centinaia di estrazioni per raggiungere un elevato livello di accuratezza nella stima dei parametri, poiché l'incertezza è confinata solamente alla frazione di informazione che è mancante nei dati. In un contesto univariato, se con *r* l'aumento relativo in varianza dovuto alle non-risposte

[9] La radice quadrata di *T* fornisce la misura dello standard error di $\hat{\theta}$. Ai fini inferenziali, Rubin (1987) suggerisce l'uso di $T^{1/2}\left(\hat{\theta}-\theta\right)$ che segue una distribuzione *t* di Student con gradi di libertà pari a $v = (m-1)\left[1+W\left(1+m^{-1}\right)^{-1}B^{-1}\right]^2$. [i]

$$r = \left(1 + m^{-1}\right) B / W \qquad [5]$$

e con *v* indichiamo i gradi di libertà, allora il tasso di informazione mancante λ causato dalla presenza di unità non osservate sarà

$$\gamma = \frac{r + \frac{2}{v+3}}{r+1} \qquad [6]$$

In base a λ, il guadagno relativo in efficienza utilizzando un numero *infinito* di imputazioni piuttosto che *m* è solitamente minimo ed è quantificabile in base alla seguente relazione

$$e_{MI} = (1 + \lambda / m)^{-1} \qquad [7]$$

Nell'applicazione che segue, si vedrà che un numero di estrazioni pari a 10 è risultato più che adeguato data la percentuale relativamente contenuta di osservazioni mancanti in ogni *item*.

3 – *Applicazione*

3.1 Il dataset

In questa sezione è riportato un esempio pratico dei diversi approcci nel trattare i dati mancanti, con particolare riferimento alla Multiple Imputation. Il contesto è quello della valutazione dello strumento di misura della qualità della didattica universitaria. Oramai pratica consolidata, infatti, ogni Ateneo pubblico italiano elabora e somministra alla fine di ogni semestre un proprio questionario mirato a raccogliere le opinioni degli studenti frequentati riguardo la didattica universitaria. L'Ateneo cui si fa riferimento è quello di Palermo, utilizzando i dati relativi alla somministrazione avvenuta presso la Facoltà di Economia durante l'ultima settimana del primo semestre dell'anno accademico 2004-2005: sono stati raccolti 2843 questionari relativi a 86 insegnamenti differenti previsti per i Corsi di Laurea triennali. Il questionario proposto si articola in diverse sezioni, relative alla raccolta delle informazioni generali sul corso e delle caratteristiche socio-demografiche del rispondente, dell'opinione riguardo all'insegnamento, all'interesse, alla soddisfazione e all'organizzazione generale, alle infrastrutture e al responsabile dell'insegnamento. Infine vi è una sezione relativa alla valutazione del 'modulo' didattico, laddove previsto. Nel caso della Facoltà di Economia, questo non sarà preso in considerazione, poiché non vi sono insegnamenti organizzati in moduli. Inoltre, nell'analisi che segue, alcuni *item* non sono presi in esame a causa della loro struttura che non consente una misurazione su scala almeno ordinale, o perché non inerenti alla valutazione dell'attività didattica in senso stretto. Pertanto, gli *item* considerati nell'analisi, dai 25 iniziali, sono stati ridotti ad un numero pari a 13 (tab. 1).

Da una breve analisi dei dati mancanti, si nota che questi sono presenti in tutti gli *item*: le unità non osservate, nella maggior parte dei casi, superano il 4% per *item* raggiungendo un massimo dell'11,2% per l'*item* F04. Se si volesse intraprendere l'analisi dei soli casi completi si dovrebbe eliminare circa il 37,6% dei record. L'assunzione di distribuzione MCAR non sembra essere valida se si sposta l'attenzione alla distribuzione dei *missing value* per ciascun record, condizionatamente alle caratteristiche personali degli studenti che hanno compilato il questionario: alcuni fattori potrebbero essere legati alle cause che hanno determinato la mancata risposta.

Tabella 1 – Elenco degli item considerati nell'analisi

Item	Dimensione indagata
B03	Chiarezza circa gli obiettivi
B04	Chiarezza circa le modalità esame
B05	Sovrapposizione contenuti
B08	Adeguatezza materiale
B10	Proporzione del carico di lavoro con CFU
B11	Coordinamento contenuti
C02	Soddisfazione
F01	% ore svolte da docente
F03	Rispetto orario lezione
F04	Rispetto orario ricevimento
F05	Disponibilità chiarimenti
F06	Stimolazione interesse
F07	Chiarezza espositiva

In particolare, la probabilità di avere un dato mancante sembra dipendere essenzialmente da due variabili: l'*item* B01 (% di ore di lezione frequentate) e l'*item* A08 (condizione occupazionale dello studente). La percentuale di almeno un *missing value* negli *item* è pari al 47,5% per quegli studenti che hanno frequentato al più il 50% delle ore di lezione, mentre per coloro che hanno frequentato più del 75% delle ore di lezione la percentuale si riduce al 33,7%. In relazione allo stato occupazionale, si rileva il 33,7% di *missing* tra coloro che non lavorano, contro il 40,3% di coloro che svolgono un impiego anche *part-time*.

Un elemento fondamentale in qualsiasi analisi statistica è l'identificabilità del dato. Un problema che si può presentare, in questo senso, nell'analisi dei questionari compilati dagli studenti frequentanti è che uno stesso studente venga interpellato per esprimere un giudizio su più corsi: nulla vieta allo studente di frequentare più insegnamenti durante un semestre e quindi può accadere che questo venga coinvolto in più di una rilevazione. Poiché l'analisi non viene svolta separatamente per ogni insegnamento, bensì a livello di Tipo di Laurea (in questo caso Laurea triennale), è plausibile che più di un questionario sia stato compilato dallo stesso soggetto, minando il principio di indipendenza tra le osservazioni. Pertanto il dataset iniziale di 2843 questionari raccolti è stato ridotto in modo tale da assicurare la corrispondenza univoca tra compilatore e questionario: in base alle informazioni circa il rispondente raccolte durante la rilevazione, si è potuto tracciare l'insieme dei profili dei rispondenti; nel caso in cui ad un profilo fossero associati più questionari, si è provveduto ad estrarre casualmente uno di essi, eliminando i rimanenti. Così si è pervenuti ad un insieme di 1605 questionari associati ad altrettanti profili. Tale popolazione, così individuata, può essere considerata una delle possibili "realizzazioni campionarie" della superpopolazione dei frequentanti, a noi ignota, ma che ipotizziamo sia costituita dagli studenti individuati dai profili.

In tale campione vi sono 384 record che presentano almeno un dato mancante, per un totale di 1105 *missing data*: da sottolineare che la percentuale più alta di dati mancanti si concentra nella sezione relativa alla valutazione del titolare dell'insegnamento, per la quale si rileva, mediamente, il 7% di dati mancanti per *item*.

In tale contesto, dove più *item* misurano lo stesso costrutto "*qualità della didattica"* e, quindi, si osservano misurazioni ripetute sullo stesso soggetto, si ritiene che la struttura dei dati osservata possa aiutare nel per predire i valori mancanti in un *item* in base a ciò che si è osservato negli altri.

3.2 – Il modello di riferimento

Seguendo il percorso già tracciato (Boscaino, 2005), si vuole valutare l'effetto dei diversi approcci al trattamento dei dati mancanti sull'efficienza delle stime dei parametri del modello ELM di Rasch, adottato per la 'taratura' del questionario.

Il modello di Rasch è un modello unidimensionale che si fonda su due aspetti fondamentali:

- più un *item* è associato ad alti livelli di qualità del servizio, più è verosimile che un soggetto esprima soddisfazione per quell'aspetto;
- più il soggetto è soddisfatto dell'intero servizio, più è verosimile che giudichi positivamente l'*item* rispetto ad un soggetto meno soddisfatto.

Il modello di Rasch assume che la probabilità che un dato soggetto si esprima in favore di un *item* è una funzione logistica della distanza tra il parametro di posizione dell'*item* e il parametro di posizione del soggetto lungo lo stesso *continuum* o tratto latente che si vuole misurare. Il pattern atteso di risposte all'insieme di *item* è determinato in base alla stime dei due parametri. Quando il pattern osservato coincide o non si discosta troppo da quello atteso, gli *item* adattano il modello di misura e costituiscono una scala di Rasch. In tal caso, il modello di Rasch trasforma le misure su scala ordinale a misure su scala ad intervalli logit.

Inserito nell'ambito dell'*Item Response Theory*, l'*Extended Logistic Model* di Rasch (Andrich 1985 e 1988), in un contesto di *item* politomici con categorie di risposta ordinate, considera la probabilità che un soggetto *n* fornisca una determinata risposta (x_{ni}) all'*i*-esimo *item* come funzione di tre parametri: β_n relativo alla soddisfazione del soggetto, δ_i relativo al livello di qualità espressa dall'*item* e τ_{ik} che attiene alla probabilità di rispondere a una delle *k* (*k=1, ...,* m) categorie di risposta previste per l'*item* considerato[10]:

$$Pr\left(X = x_{ni}\right) = \frac{exp\left[\tau_{i_x} + x_{ni}\left(\beta_n - \delta_i\right)\right]}{\sum_{k=1}^{m} exp\left[\tau_{i_k} + k\left(\beta_n - \delta_i\right)\right]}. \qquad [8]$$

In breve[12], il modello consente il confronto diretto tra i parametri relativi ai soggetti e quelli relativi agli *item*, al fine di misurare un concetto latente rappresentato da un unico *continuum* unidimensionale. Ciascun parametro di posizione dell'*item* stimato lungo il *continuum* concorre alla spiegazione di una "porzione" del concetto in oggetto d'esame. Il confronto con le posizioni stimate dei parametri relativi ai soggetti lungo lo stesso *continuum*[11] consente, inoltre, di valutare l'adeguatezza dell'organizzazione del questionario e il grado di soddisfazione (nel nostro esempio) che questo è riuscito a misurare.

Per quanto riguarda la verifica dell'adattamento dei dati al modello, le statistiche inerenti la bontà delle stime dei parametri relativi agli *item*, note in letteratura, sono l'*outfit* e *infit mean square* (Bond e Fox, 2001). "*Outfit mean square*" è un modo breve per "*Outlier sensitive mean square residual goodness of fit statistic*". L'*outfit* misura la discrepanza tra i dati e il modello: per ciascun soggetto si determina la somma dei quadrati dei residui degli *item* dividendola per il numero di *item* a cui il soggetto è stato sottoposto. Tale statistica, però, è sensibile ai valori estremi, ovvero alle risposte inattese fornite dai soggetti che esprimo livelli di soddisfazione "lontani" dal livello di qualità di cui l'*item* è espressione (Linacre, 2002). Un'alternativa, proposta da Wright e Masters (1982), è l'*infit mean square* (*information weighted mean square residual goodness of fit statistic*), calcolata attribuendo a ciascun residuo un peso pari alla sua varianza. Dividendo tale risultato per la somma delle varianze, si ottiene la stessa distribuzione degli *outfit* prendendo però in considerazione i diversi pesi.

[10] Per una più completa trattazione del modello di Rasch, si rimanda a: Rasch G. (1960), *Probabilistic Models for some Intelligence and attainment tests*, Copenhagen Danish Institute for Educational Research; Andrich D. (1988), *Rasch models for measurement*, Sage, Beverly Hills; Fischer G.H., Molenaar I. (1995), *Rasch Models Foundations, Recent Developments, and Applications*, Springer, NewYork.

[11] In base ai punteggi ottenuti dai soggetti e per gli *item*, si dimostra (Fischer e Molenaar, 1995) che tale modello consente una trasformazione monotona di scala da dati a livello ordinale a dati a livello intervallare, sia per i soggetti che per gli *item*, attribuendo alle stime la medesima unità di misura (il *logit*), rendendo possibile il confronto diretto tra soddisfazione del soggetto e qualità dell'*item*, in termini di distanze delle posizioni stimate lungo lo stesso *continuum* considerato (vedi nota 12).

Per tale ragione i ricercatori spesso considerano maggiormente i punteggi *infit* che quelli *outfit*, anche se l'*infit* risulta essere influenzato dai *pattern* di rispostaStatisticamente, i *mnsq* sono statistiche X^2 divise per i propri gradi di libertà; il valore atteso è pari ad 1; valori maggiori di 1 indicano una condizione di *underfit,* ovvero scarso adattamento dei dati al modello, mentre la situazione opposta è indicatrice *overfit*, ovvero di ridondanza dei dati (Linacre, 2002). L'interpretazione di quando il *mean square* sia troppo grande o troppo piccolo é dettata dall'esperienza e dal contesto in cui si opera. Wright e Linacre (1994) forniscono una guida pratica per un ragionevole range dell'*item mean square infit* e *outfit* (tab. 2) a seconda dell'ambito in cui si opera: in un contesto quale il nostro, l'intervallo 0,4 – 1,2 è preso come riferimento come elemento discriminatorio nella fase di selezione degli *item*.

Tabella 2 – Intervalli di *item mean square* ragionevoli per *infit* e *outfit*

Type of test	Range
Multiple-choice test (High stakes)	0,8 – 1,2
Multiple-choice test (Run of the mill)	0,7 – 1,3
Rating Scale (Likert/survey)	0,6 – 1,4
Clinical observation	0,5 – 1,7
Judged (where agreement is encouraged)	0,4 – 1,2

3.3 Il confronto tra i metodi

Il dataset dei 1605 questionari, come già accennato, è stato trattato secondo 4 approcci differenti alla gestione dei dati mancanti. Pertanto, al dataset risultante da ognuno dei 4 metodi si è applicato il modello ELM di Rasch tramite l'utilizzo del software Winsteps 3.61.

Per quanto riguarda l'imputazione multipla (MI), come già anticipato, questa è stata ripetuta per 10 volte: la Tabella 3 riporta le stime dei parametri di posizione degli *item* (e dei relativi standard error) ottenuti per i 10 campioni generati dal processo di imputazione multipla. Il confronto tra i 4 differenti metodi è possibile dopo che si è operata, tramite la [1] e la [4], la sintesi delle stime dei parametri relativi agli *item* e dei corrispondenti standard error ottenuti in base al MI. La Tabella 4 riporta le stime (e gli standard error) dei parametri in base ai campioni utilizzati: il campione dei soli dati completi (a), le stime di sintesi dei dieci campioni ottenuti dalla procedura di imputazione multipla (MI) (d), il campione in cui i dati mancanti sono stati sostituiti dal valore modale dell'*item* (b) e, infine, il campione ottenuto con il metodo di regressioni sequenziali (c). La Tabella 4 consente di effettuare il confronto, in termini di guadagno in efficienza, tra i diversi metodi considerati: dalla Tabella 5, infatti, si evince che, confrontando gli standard error delle stime ottenute con le 4 tecniche, rispetto all'analisi dei soli dati completi (a), il metodo di imputazione multipla (MI) riduce lo standard error del 14% (valore medio). Un guadagno in precisione (4,15% in media) si ottiene pure rispetto al metodo regressivo (c), mentre rispetto alla sostituzione del dato mancante con il valore modale dell'*item* (b), la precisione si riduce mediamente del 4,41%. Da tali risultati emerge un chiaro vantaggio nell'adottare il MI in termini di precisione delle stime[12]. Il risultato ottenuto nel caso della sostituzione dei *missing data* con il valore modale era, comunque, atteso: infatti, tale sostituzione si è ripercossa sulla variabilità dei dati che, in tale situazione, si è ridotta più che negli altri casi.

[12] Il risultato è avvallato anche dai risultati di simulazione ottenuti da Sulis (*A multiple imputation approach to correct non response bias in an analysis of student ratings of university courses,* in tesi di dottorato 2007), dove emerge che la MI ricostruisce in maniera abbastanza accurata sia i valori delle stime dei parametri degli *item* che i loro standard error.

Tabella 3 – Stime dei parametri di posizione degli *item* per i 10 campioni imputati (d)

ITEM	C01		C02		C03		C04		C05		C06		C07		C08		C09		C10	
	Stima	SE	Stima	SE	Stima	SE	Stima	SE	Stima	SE	Stima	SE	Stima	SE	Stima	SE	Stima	SE	Stima	SE
B03 – Chiarezza riguardo gli obiettivi	0,01	0,03	0,01	0,03	0,01	0,03	0,01	0,03	0,00	0,03	0,00	0,03	0,00	0,03	0,02	0,03	0,00	0,03	0,01	0,03
B04 – Chiarezza delle modalità esame	0,30	0,03	0,30	0,03	0,30	0,03	0,29	0,03	0,29	0,03	0,29	0,03	0,29	0,03	0,30	0,03	0,30	0,03	0,30	0,03
B05 – Sovrapposizione contenuti	0,92	0,03	0,91	0,03	0,92	0,03	0,92	0,03	0,91	0,03	0,93	0,03	0,92	0,03	0,92	0,03	0,90	0,03	0,92	0,03
B08 – Adeguatezza materiale	0,26	0,03	0,25	0,03	0,26	0,03	0,25	0,03	0,24	0,03	0,25	0,03	0,25	0,03	0,25	0,03	0,26	0,03	0,24	0,03
B10 – Prop. carico-CFU	0,62	0,03	0,61	0,03	0,61	0,03	0,61	0,03	0,62	0,03	0,61	0,03	0,61	0,03	0,62	0,03	0,62	0,03	0,63	0,03
B11 – Coordinamento contenuti	0,58	0,03	0,58	0,03	0,58	0,03	0,58	0,03	0,58	0,03	0,58	0,03	0,57	0,03	0,57	0,03	0,56	0,03	0,57	0,03
C02 – Soddisfazione	0,17	0,03	0,16	0,03	0,17	0,03	0,18	0,03	0,17	0,03	0,17	0,03	0,17	0,03	0,17	0,03	0,17	0,03	0,17	0,03
F01 – % ore svolte da docente	-0,88	0,05	-0,89	0,05	-0,88	0,05	-0,87	0,05	-0,89	0,05	-0,89	0,05	-0,87	0,05	-0,88	0,05	-0,88	0,05	-0,89	0,05
F03 – Rispetto orario lezione	-0,54	0,04	-0,52	0,04	-0,53	0,04	-0,53	0,04	-0,54	0,04	-0,56	0,04	-0,55	0,04	-0,56	0,04	-0,56	0,04	-0,56	0,04
F04 – Rispetto orario ricevimento	-0,56	0,04	-0,51	0,04	-0,53	0,04	-0,55	0,04	-0,54	0,04	-0,53	0,04	-0,52	0,04	-0,57	0,04	-0,53	0,04	-0,52	0,04
F05 – Disponibilità chiarimenti	-0,86	0,04	-0,87	0,04	-0,86	0,04	-0,87	0,04	-0,83	0,04	-0,85	0,04	-0,85	0,04	-0,84	0,04	-0,84	0,04	-0,87	0,04
F06 – Stimolazione interesse	0,03	0,03	0,02	0,03	0,01	0,03	0,04	0,03	0,02	0,03	0,04	0,03	0,02	0,03	0,03	0,03	0,03	0,03	0,03	0,03
F07 – Chiarezza espositiva	-0,05	0,03	-0,05	0,03	-0,06	0,03	-0,05	0,03	-0,05	0,03	-0,03	0,03	-0,05	0,03	-0,03	0,03	-0,04	0,03	-0,05	0,03

Tabella 4 – Stima dei parametri di posizione degli *item* secondo i diversi approcci

ITEM	Completi (a)		MI (d)		Moda (b)		Regressivo (c)	
	Stima	SE	Stima	SE	Stima	SE	Stima	SE
B03 – Chiarezza obiettivi	0,00	0,04	0,01	0,03	0,05	0,03	0,01	0,03
B04 – Chiarezza modalità esame	0,31	0,04	0,30	0,03	0,33	0,03	0,31	0,03
B05 – Sovrapposizione contenuti	0,97	0,03	0,92	0,03	0,93	0,03	0,94	0,03
B08 – Adeguatezza materiale	0,25	0,04	0,25	0,03	0,28	0,03	0,27	0,03
B10 – Prop. carico-CFU	0,61	0,04	0,62	0,03	0,65	0,03	0,64	0,03
B11 – Coordinamento contenuti	0,62	0,04	0,58	0,03	0,60	0,03	0,59	0,03
C02 – Soddisfazione	0,16	0,04	0,17	0,03	0,20	0,03	0,18	0,03
F01 – % ore svolte da docente	-0,90	0,06	-0,88	0,05	-0,91	0,05	-0,94	0,05
F03 – Rispetto orario lezione	-0,55	0,04	-0,55	0,04	-0,57	0,04	-0,59	0,04
F04 – Rispetto orario ricevimento	-0,59	0,04	-0,54	0,04	-0,62	0,04	-0,56	0,04
F05 – Disponibilità chiarimenti	-0,88	0,05	-0,86	0,04	-0,87	0,04	-0,88	0,04
F06 – Stimolazione interesse	0,05	0,04	0,03	0,03	0,05	0,03	0,06	0,03
F07 – Chiarezza	-0,04	0,04	-0,05	0,03	-0,12	0,03	-0,02	0,03

Tabella 5 –Variazione percentuale dello SE delle stime confrontando MI con gli altri metodi

ITEM	MI *vs* a	MI *vs* b	MI *vs* c
B03 – Chiarezza obiettivi	-22,94%	2,75%	-2,67%
B04 – Chiarezza modalità esame	-23,79%	1,62%	-1,59%
B05 – Sovrapposizione contenuti	4,06%	4,06%	-3,90%
B08 – Adeguatezza materiale	-22,54%	3,27%	-3,17%
B10 – Prop. carico-CFU	-22,79%	2,94%	-2,86%
B11 – Coordinamento contenuti	-22,74%	3,01%	-2,92%
C02 – Soddisfazione	-23,99%	1,35%	-1,33%
F01 – % ore svolte da docente	-15,53%	1,36%	-1,34%
F03 – Rispetto orario lezione	7,55%	7,55%	-7,02%
F04 – Rispetto orario ricevimento	11,69%	11,69%	-10,47%
F05 – Disponibilità chiarimenti	-14,56%	6,80%	-6,36%
F06 – Stimolazione interesse	-20,98%	5,36%	-5,08%
F07 – Chiarezza	-20,84%	5,55%	-5,26%
Media	-14,42%	4,41%	-4,15%

In base alla [7], in Tabella 6 sono riportati i valori di e_{MI} per le stime dei parametri di posizione degli *item* ottenute con il metodo MI: si può facilmente apprezzare che la scelta di replicare 10 volte l'imputazione è stata più che idonea allo scopo.

Tabella 6 – Efficienza delle stime dei parametri di posizione degli *item*, ottenute mediante MI

ITEM	e_{MI}
B03 – Chiarezza obiettivi	0,9995
B04 – Chiarezza modalità esame	1,0000
B05 – Sovrapposizione contenuti	1,0000
B08 – Adeguatezza materiale	1,0000
B10 – Prop. carico-CFU	1,0000
B11 – Coordinamento contenuti	1,0000
C02 – Soddisfazione	1,0000
F01 – % ore svolte da docente	1,0000
F03 – Rispetto orario lezione	1,0000
F04 – Rispetto orario ricevimento	1,0000
F05 – Disponibilità chiarimenti	1,0000
F06 – Stimolazione interesse	0,9997
F07 – Chiarezza	1,0000

3.4 La ‘taratura’ del questionario

La fase di ‘taratura’ del questionario, ovvero di individuazione del migliore set di *item* per misurare il concetto di qualità della didattica, si basa su un processo *backward selection* degli *item* fondata sulle statistiche di adattamento dei dati al modello di Rasch e, soprattutto, su considerazioni dettate dal buon senso: l’individuazione di un ‘buon’ modello, svolta esclusivamente sulla base di un test di bontà di adattamento, può risultare limitante se non errata.

Il procedimento iterativo è stato svolto su tutti i 13 campioni: il modello di Rasch è stato applicato sul campione dei soli dati completi, sui dieci campioni generati dalla imputazione multipla (MI), sul campione in cui i dati mancanti sono stati sostituiti dal valore modale dell’*item* e, infine, sul campione ottenuto con il metodo regressivo.

Attraverso la procedura iterativa si è eliminato, ad ogni passo, l’*item* che mostra il più elevato *misfit*, procedendo quindi alla stima del nuovo modello. Il processo ha avuto termine col raggiungimento di un modello i cui *item* presentano statistiche di adattamento in linea con quanto suggerito da Wright e Linacre (1994) e, come sarà più chiaro successivamente, anche con opportune considerazioni svolte dal ricercatore.

I risultati ottenuti sono in linea tra i diversi campioni: l’insieme di *item* eliminati è risultato sempre lo stesso e le statistiche di adattamento riportano valori pressoché uguali. Il primo *item* eliminato, in tutti i campioni, è stato quello relativo alla sovrapposizione dei contenuti dell’insegnamento oggetto di valutazione con gli altri insegnamenti (B05). Graficamente, il *misfit* di tale *item* può essere apprezzato mediante l’*Item Characteristic Curve* (ICC), ovvero la curva che riflette la probabilità di rispondere ad una categoria dell’*item* in funzione del parametro relativo al soggetto. In Figura 1, in particolare, sono riportate le ICC relative all’*item* B05 costruite per i 13 campioni: per ognuna di esse si nota lo scarso adattamento dei dati (spezzata empirica) al modello (curva di probabilità, ogiva continua), evidenziato dall’intervallo di confidenza al 95% (i cui limiti superiore ed inferiore sono rappresentati dalle spezzate), mentre il confronto tra le curve dei diversi campioni mette in luce una sostanziale invarianza tra i risultati, in termini di adattamento, ottenuti nei diversi casi e con le differenti metodologie.

Tale situazione si mantiene anche nei passi successivi dell’iterazione (per questo motivo non sono riportati ulteriori confronti grafici): sebbene l’ordine di eliminazione degli *item* non sia sempre lo stesso per i campioni, il set di *item* eliminati alla fine del processo iterativo è invece uguale:

- B05 – Sovrapposizione contenuti tra insegnamenti;
- F01 – % ore svolte dal docente;
- B10 – Carico di studi commisurato ai crediti;
- B11 – Coordinamento contenuti tra diversi insegnamenti.

Il processo di selezione ha avuto termine col raggiungimento di un modello i cui *item* presentano statistiche di adattamento accettabili. In questa fase il buon senso del ricercatore ha assunto un ruolo fondamentale: se ci si fosse basati sull’utilizzo meccanico e cieco delle statistiche di *misfit*, il processo iterativo non avrebbe raggiunto conclusione. Si è deciso di interrompere l’iterazione al 4° passo, dato che le statistiche di *misfit* assumono valori poco al di fuori dell’intervallo di riferimento e che gli *item* rimanenti sembrano essere parte logica essenziale della struttura dello strumento di misura della qualità della didattica:

- B03 – Chiarezza obiettivi del corso;
- B04 – Chiarezza modalità esame;
- B08 – Adeguatezza materiale;
- C02 – Soddisfazione generale;
- F03 – Rispetto orario lezione da parte del docente;
- F04 – Rispetto orario ricevimento da parte del docente;

- F05 – Disponibilità del docente ai chiarimenti;
- F06 – Stimolazione dell'interesse;
- F07 – Chiarezza del docente nell'esporre gli argomenti.

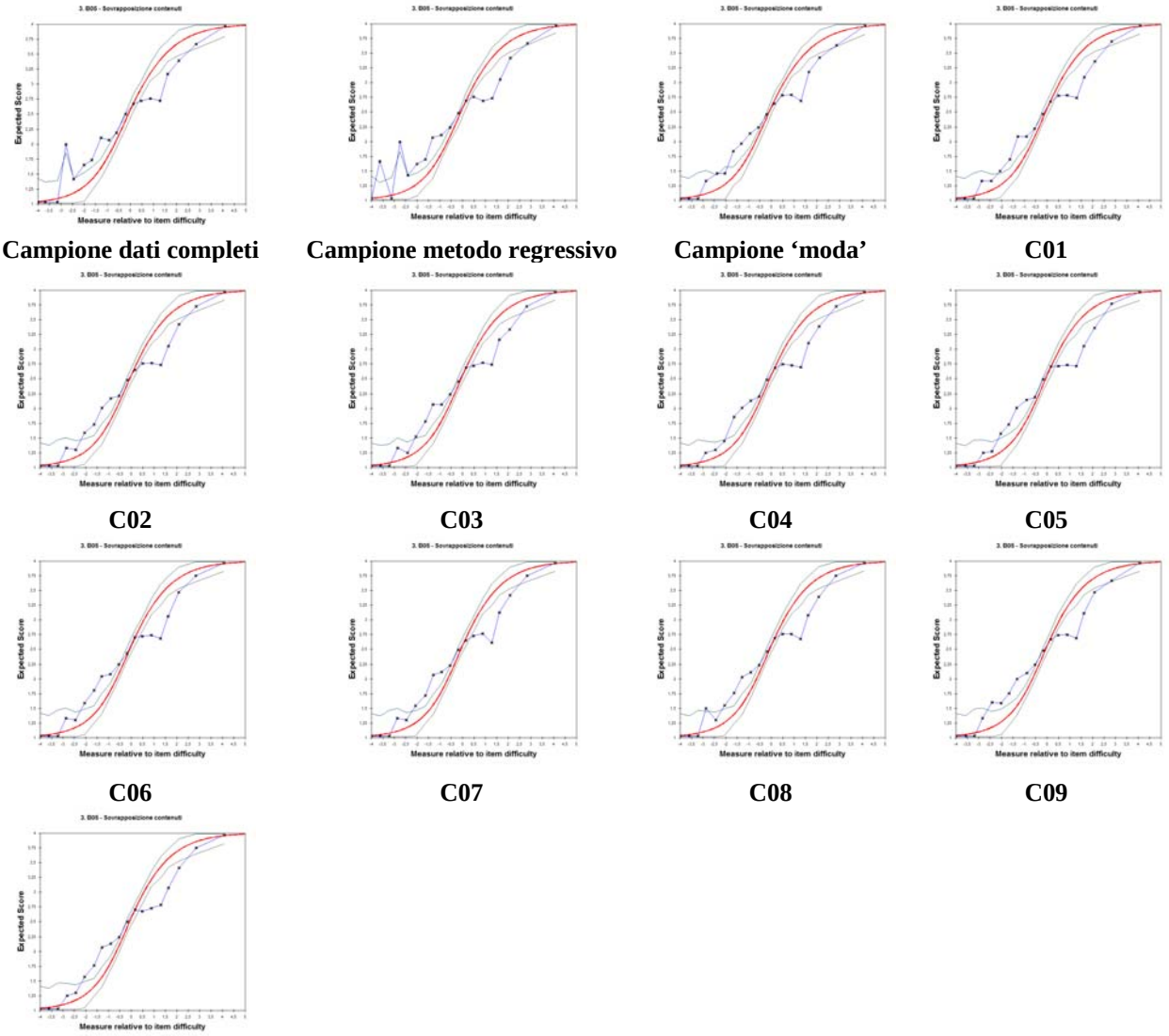

Figura 1 – ICC stimate ed empiriche e bande di confidenza per l'*item* B05 nei diversi campioni

Gli *item* eliminati dal processo valutativo del modello sottolineano differenti cause di *misfit*: il B05, B10 e B11 probabilmente attengono ad una valutazione che ha a che fare con *item* che verosimilmente richiedono una conoscenza complessiva che lo studente ancora non si è formato; l'*item* F01 poco si adatta ad un contesto come quello della Facoltà di Economia di Palermo in cui i docenti titolari degli insegnamenti spesso sono affiancati e/o coadiuvati da uno o più docenti (tra titolari di assegni di ricerca e ricercatori) ai quali è affidata parte del programma previsto.

Gli *item* rimasti, invece, sono relativi a caratteristiche peculiari dell'attività didattica: chiarezza del docente, adeguatezza del materiale didattico fornito e consigliato, rispetto degli orari e disponibilità del docente a fornire chiarimenti durante le lezioni, oltre al grado di soddisfazione generale per l'insegnamento.

La *bubble chart* riportata nel Grafico 1 (relativo, per brevità, al solo campione C01), costruita in base alle stime delle posizioni dei parametri degli *item* sul *continuum* (*measures*, asse delle ordinate), ai corrispettivi standard error (il raggio delle circonferenze) e alle misure di *infit* relative ai

singoli *item* (asse delle ascisse), mette in evidenza la sovrapposizione delle stime lungo il *continuum*, a cui potrebbe corrispondere, talvolta, la sovrapposizione concettuale del livello di qualità espresso del concetto latente.

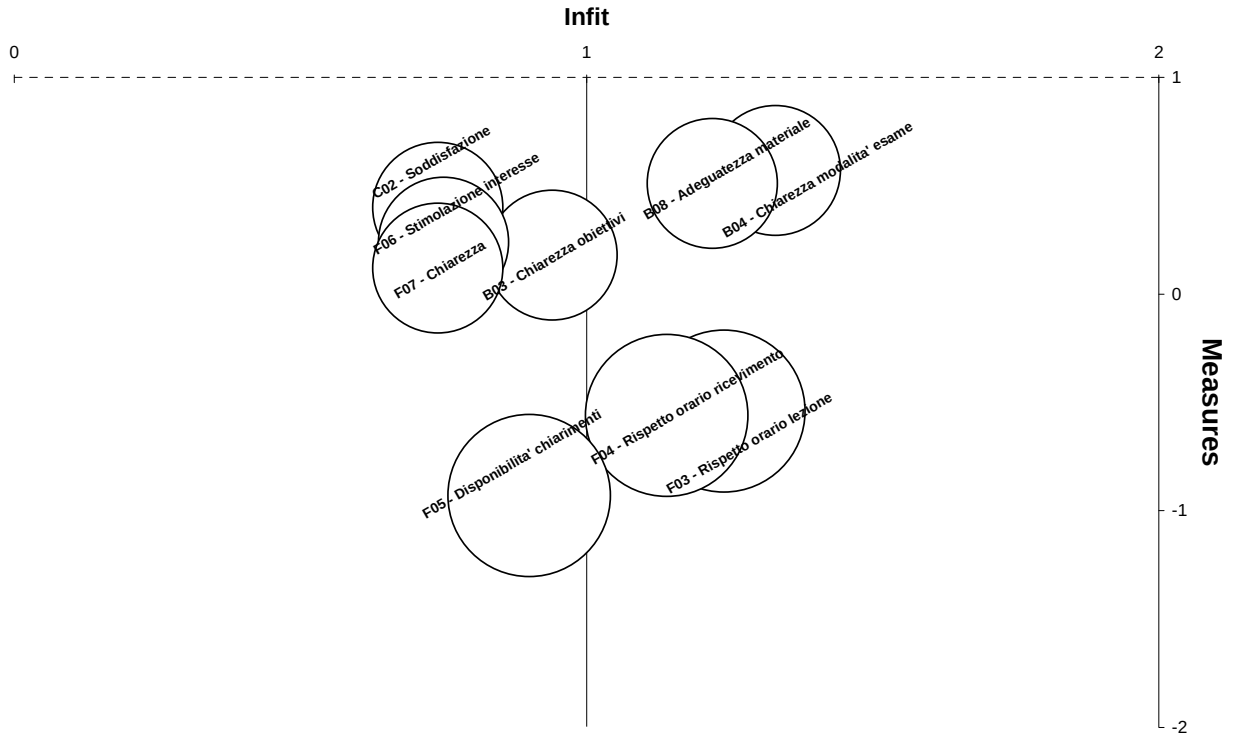

Grafico 1 – *Bubble chart* relativo al set ottimo di *item* stimato per il campione C01

Considerazioni conclusive

La gestione dei dati mancanti è un problema che, sebbene rilevante, non sempre è affrontato adeguatamente. Molti ricercatori svolgono una semplice analisi dei dati completi assumendo una distribuzione MCAR dei dati mancanti, supportati anche dai software statistici che in molti casi implementano tale funzione. Il rischio di distorsione e di perdita di efficienza (oltre la notevole riduzione del set di dati, con la conseguente diminuzione del potere statistico in fase inferenziale) è quindi sottovalutato. Nell'ambito dell'*Item Response Theory*, inoltre, l'importanza di un dataset completo, in fase di stima dei parametri, è cruciale. La tecnica di imputazione multipla basata sulle distribuzioni condizionate si è rilevata adeguata al contesto della valutazione della didattica. La procedura ha infatti permesso sia di utilizzare a fine predittivi l'ampia informazione proveniente dalla presenza di misurazioni ripetute sulla stessa unità sia di tenere espressamente in considerazione l'incertezza legata al vero valore del dato mancante. Ciò ha consentito di limitare lo "spreco" di dati raccolti, nonché di ottenere stime dei parametri più efficienti rispetto agli altri metodi.

Il principale vantaggio rispetto al procedere con un'analisi delle osservazioni complete è identificabile, oltre che, ovviamente, nel mantenimento della dimensione campionaria, nel non avere assunto, a priori, che l'informazione proveniente dalle unità parzialmente osservate fosse ignorabile. Assunzione, questa, che dai risultati delle prime analisi esplorative sembrava essere violata.

La calibrazione del questionario ha condotto verso l'individuazione di un set ottimo di *item* identico per tutti i campioni considerati con le diverse metodologie di approccio ai dati mancanti.

Il processo iterativo è fortemente condizionato dalle scelte di 'buon senso', ovvero supportate dalle statistiche di adattamento ma non esclusivamente da quelle. Rasch stesso ha suggerito l'uso di statistiche di adattamento basate sull'X^2 per determinare quanto bene un set di dati empirici si attenga agli assiomi e alle caratteristiche del proprio modello, ma come Wright e Linacre (1994) fanno notare,

quando un *mean square* è troppo grande o troppo piccolo? Non esistono regole chiare e semplici in merito: solo la pratica, l'esperienza e il contesto in cui si opera possono essere da guida. Se ci soffermiamo un attimo a pensare a tutti i limiti della statistica X^2, ad esempio, ci si può rendere conto della problematica che si sta affrontando. In realtà il problema potrebbe stare nell'ipotesi nulla che poniamo: i dati adattano il modello? Non siamo interessati all'adattamento perfetto, che non è mai ottenibile dai dati empirici. Affidarsi unicamente alle statistiche di adattamento non può rientrare nel *modus operandi* di un qualsiasi ricercatore: l'idea che il buon senso debba giocare un ruolo sempre maggiore, nel supportare il ricercatore in fase di determinazione del modello di analisi, non deve essere scartata, soprattutto nell'ambito della ricerca nelle scienze umane in cui qualsiasi tratto latente diventa sempre più 'intangibile' e la componente di errore assume caratteristiche sempre nuove e diverse.

Bibliografia di riferimento

Andrich D. (1985), *An elaboration of Guttman scaling with Rasch models for measurement*, in Tuma N. B. (A cura di), *Sociological methodology*, Jossey-Bass, San Francisco CA.

Andrich D. (1988), *Rasch models for measurement*, Sage, Beverly Hills.

Bond T.G. e Fox C.M. (2001), *Applying the Rasch Model – Fundamental Measurement in the Human Sciences*, Erlbaum, NJ

Boscaino G. (2005), *La qualità della didattica: la calibrazione dello strumento di misura con il modello di Rasch*, Tesi di dottorato in Statistica Applicata.

Fischer G.H. e Molenaar I. (1995), *Rasch Models Foundations, Recent Developments, and Applications*, Springer, NewYork.

Glasser M. (1964), Linear regression analysis with missing observations among the independent variables, *Journal of the American Statistical Association*, n. 59.

Haitovsky Y. (1968), Missing data in regression analysis, *Journal of the Royal Statistical Society (B)*, n. 30.

Knapp T. (1998), *Quantitative nursing research*, Sage Publications, Thousand Oaks, CA.

Linacre J.M. (2002), *What do Infit and Outfit Mean-Square and Standardized mean?*, Rasch Measurement Transaction, n.16

Little R.J.A. (1992), Regression with missing X's: a review, *Journal of the American Statistical Association*, n. 87.

Little R.J.A. e Rubin D.B. (2002), *Statistical Analysis with Missing Data*, John Wiley, New York.

Masters G.N. e Wright B.D. (1982), *Rating scale analysis: Rasch measurement*, MESA Press, Chicago.

Raghunathan T. E. (2004), What Do We Do With Missing Data? Some Options for Analysis of Incomplete Data, *Annual Review Public Health* , n.25.

Raghunathan T. E., Lepkowsky J.M., van Hoewyk M. e Solenberger P.W. (2001), A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, *Survey Methodology*, n. 27.

Raymond M.R. (1986), Missing data in evaluation research, *Evaluation and the Health Profession*, n. 9.

Rasch G. (1960), *Probabilistic Models for some Intelligence and attainment tests*, Copenhagen Danish Institute for Educational Research.

Rubin D.B. (1976), Inference and missing data, *Biometrika*, n. 63.

Rubin D.B. (1987), *Multiple imputation for nonresponse in survey*, John Wiley, New York.

Rubin D.B. (1996), Multiple imputation after 18 years; *Journal of the American Statistical Association*, n. 91.

Schafer J.L. (1997), *Analysis of incomplete multivariate data*, Chapman and Hall, Londra.

Schafer J.L. e Graham J.W. (2002), Missing data: our view of the state of the art, *Psychological Methods*, vol. 7, n. 2.

Stone P.W. (2001), What's the big deal about missing data, *Applied nursing research*, vol.14, n.4.

Sulis I. (2007), An analysis of missing data structure: a simulation study to assess the loss in term of accuracy in the estimation of item parameters, in *Measuring students' assessments of "university course quality" using mixed-effects models*, ,Tesi di dottorato in Statistica Applicata.

Tsikriktsis N. (2005), A review of techniques for treating missing data in OM survey research, Journal of Operation Management, n. 24, pagg. 53-62

Wright B. D., Linacre J. M. (1994), *Reasonable mean-square fit values*, Rasch Measurement Transaction, 8(3), 370, http://www.rasch.org/rmt/rmt83.htm